Overview of the 2020 WOSP 3C Citation Context Classification Task

Overview of the 2020 WOSP 3C Citation Context Classification Task

Suchetha N. Kunnath
KMi, The Open University
Milton Keynes
UK
snk56@open.ac.uk

David Pride
KMi, The Open University
Milton Keynes
UK
david.pride@open.ac.uk

Bikash Gyawali
KMi, The Open University
Milton Keynes
UK
bikash.gyawali@open.ac.uk

Petr Knoth
KMi, The Open University
Milton Keynes
UK
petr.knoth@open.ac.uk

Abstract

The 3C Citation Context Classification task is the first shared task addressing citation context classification. The two subtasks, A and B, associated with this shared task, involves the classification of citations based on their purpose and influence, respectively. Both tasks use a portion of the new ACT dataset, developed by the researchers at The Open University, UK. The tasks were hosted on Kaggle, and the participated systems were evaluated using the macro f-score. Three teams participated in subtask A and four teams participated in subtask B. The best performing systems obtained an overall score of 0.2056 for subtask A and 0.5556 for subtask B, outperforming the simple majority class baseline models, which scored 0.11489 and 0.32249, respectively. In this paper we provide a report specifying the shared task, the dataset used, a short description of the participating systems and the final results obtained by the teams based on the evaluation criteria. The shared task has been organised as part of the 8th International Workshop on Mining Scientific Publications (WOSP 2020) workshop.

1 Introduction

Citation analysis for research evaluation has been a subject of interest for the past several decades. The conventional one dimensional perspective of citation analysis, based on the pure citation frequency, which treats all citations equally, has endured a lot of criticism way back [Moravcsik and Murugesan, 1975, Kaplan, 1965]. Subsequently, researchers have emphasised the need for developing new methods that consider the different aspects of the citing sentences. One such qualitative way for measuring the scientific impact is to analyse the citation context for discovering the author's reason for citing a paper. The text containing the reference to the cited document, the citation context, has proved to be a valuable signal for characterising the citation intent [Teufel et al., 2006]. The increase in the accessibility of the scientific publications, as well as the availability of full text of the research documents, from various services like CORE [Knoth and Zdrahal, 2012] facilitates the possibility of exploring citation contexts, thereby further extending the bibliometric studies for research assessment [Pride and Knoth, 2017].

Understanding the intent of citation has an essential role in measuring the scientific impact of the research papers. The possibility of knowing why a citation is included in one's work and how influential it is offers an excellent measure for evaluating the impact of a scientific publication. Previous approaches for citation context classification employed a variety of annotation schemes ranging from low to high granularity. Due to the lack of standard methods and annotation schemes, a comparison of the earlier systems is practically difficult. Earlier systems used datasets with very limited size and this is probably because of the difficulties in manually annotating the citation contexts. Besides, most of the research on citation context classification is not extensive enough and mainly reduced to specific domains of application, for instance, computer science and biomedical fields. This raises questions related to the generalisability of the presented models.

The 3C Shared task aims to create a platform encouraging researchers to participate in research in this area so that we can more reliably measure the performance of methods that have been tried in this area, establish the state-of-the-art and understand what works and what doesn't. Two subtasks associ-

ated with this shared task provide the participating teams the possibility to explore the new Academic Citation Typing (ACT) dataset [Pride et al., 2019, Pride and Knoth, 2020] for analysing the citation context and classify the associated citations based on their purpose (subtask A) and influence (subtask B). A total of four teams participated in subtask A, and five teams participated in the subtask B. We used Kaggle InClass competitions[1] for organising this shared task and the participating systems were evaluated using the macro f-score.

This overview paper presents the 2020 3C Shared Task organisation. Section 2 describes the related work; Section 3 discusses the shared task setup, the data used, the baselines, followed by task evaluation in Section 4. Section 5 summarises the participating system description. Section 6 and 7 presents the results and the conclusion.

2 Related Work

Several supervised machine learning based frameworks that inspect the language used in scientific discourse have been developed in the past to categorise citations based on their context. [Teufel et al., 2006] used an annotation scheme with 12 categories and applied machine learning techniques on 2,829 citation contexts from 116 articles, using linguistic features including the cue phrases. These 12 classes belonged to four top-level categories; citations explicitly mentioning weakness, citations that compares or contrasts, citations which agrees or uses or is compatible with the citing work and finally a neural class. A more fine-grained classification scheme introduced by Jurgens et.al [Jurgens et al., 2018] contains six categories and 1,941 instances from papers in Computational Linguistics(ACL-ARC dataset). The authors applied three novel features: pattern-based, topic-based and prototypical argument-based features besides the structural, lexical and grammatical, field and usage features.

The above mentioned approaches all used hand-engineered features for classification. [Cohan et al., 2019] proposed a neural multi-task learning method using non-contextualised (GloVe) and contextualised word embeddings (ELMo) along with BiLSTM and attention mechanism for citation intent classification. To achieve multi-task learning, the authors used two auxiliary tasks to aid the main classification task. The new dataset (SciCite) [Cohan et al., 2019] contains 11,020 instances belonging to Computer Science and Medicine domains and only three citation categories. A pre-trained model using 1.14M papers from Semantic Scholar[2], called SciBERT [Beltagy et al., 2019], was released in 2019 and achieved a macro f-score of nearly 85% with fine-tuning using the SciCite dataset.

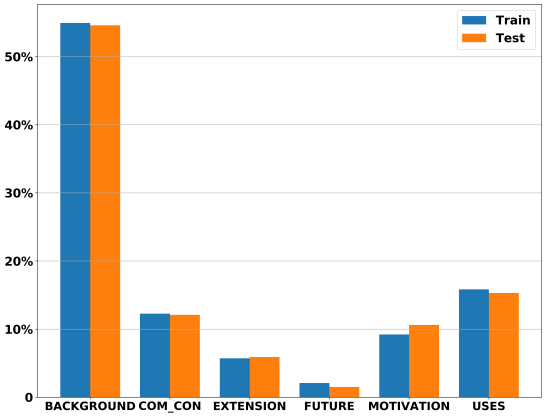

Figure 1: Subtask A data distribution

3 The 3C Shared Task

To address the limitations of citation context classification from the previous studies, we introduce a unified task to compare different citation classification methods on the same dataset. The shared task for the citation context classification, called, the "3C Shared Task", is organised as part of the International Workshop on Mining Scientific Publications (WOSP), 2020[3], collocated with the Joint Conference on Digital Libraries (JCDL) 2020[4]. As organisers, we believe, this shared task will provide the opportunity for comparing different classification systems and help progress the state-of-the-art. The competing systems in the 3C shared task will serve as a standard benchmark for future research in this direction.

3.1 Task Definition

The 3C shared task is a classification challenge, where each citation context is categorised based on its purpose and influence. The following are the output categories associated to the two subtasks respectively.

[1] https://www.kaggle.com/c/about/inclass

[2] https://www.semanticscholar.org/

[3] https://wosp.core.ac.uk/jcdl2020/index.html

[4] https://2020.jcdl.org/

unique_id	CC10
core_id	158977742
citing_title	Ontology-Based Recommendation of Editorial Products
citing_author	Thiviyan Thanapalasingam
cited_title	Ontological user profiling in recommender systems
cited_author	Middleton
citation_context	The main advantages of these solutions are i) the ability to exploit the domain knowledge for improving the user modelling process, ii) the ability to share and reuse system knowledge, and iii) the alleviation of the cold-start and data sparsity problems [16,#AUTHOR_TAG].
citation_class_label	BACKGROUND
citation_influence_label	INCIDENTAL

Table 1: ACT data format

- **Subtask A**: Multiclass classification of citation contexts based on purpose with categories - BACKGROUND, USES, COMPARES_CONTRASTS, MOTIVATION, EXTENSION, and FUTURE.
- **Subtask B**: Binary classification of citations into INCIDENTAL or INFLUENTIAL classes, i.e. a task for identifying the importance of a citation.

The shared task was managed and evaluated using the Kaggle InClass competitions, an easy to set up, free self-service platform for hosting Data Science challenges, with notebook support for GPU and code sharing. The ability to maintain a leaderboard, which allows the participants to view results immediately after submission, built-in evaluation metrics and automated submission scoring are some of the features offered by Kaggle.

Both subtasks were organised as separate competitions in Kaggle. The shared task homepage for subtask A can be found at https://www.kaggle.com/c/3c-shared-task-purpose/. The following url correspond to the competition page for the subtask B, https://www.kaggle.com/c/3c-shared-task-influence/. The task participants were required to:

- Develop methods to classify the citations based on its purpose or influence and submit the results via Kaggle
- Document and submit their method for classifying the citations as a short paper
- Provide source code for each method

The competitions lasted 43 days, starting from May 11, 2020 till June 22, 2020.

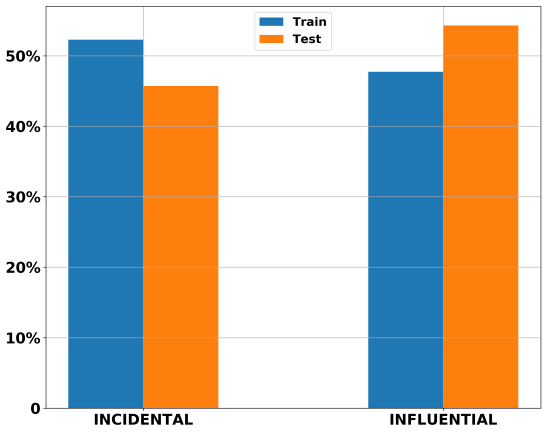

Figure 2: Subtask B data distribution

3.2 Dataset

The previous studies on citation classification systems used datasets that were annotated by domain experts and independent annotators, making the evaluation process relatively slow and expensive. Existing datasets in the field are, as a result, also confined to a specific domain, mainly computer science and biomedical domains, because this is the domain in which the annotators can could label the instances. The citation contexts need not always contain explicit signals that express the author's motivation for citing a paper. Since interpreting the citation intent is difficult for an independent annotator, authors themselves are in a better position to report their motivations in citing a paper [Pride and Knoth, 2020]. [Pride et al., 2019] used this strategy; asking authors to annotate their papers

for tagging citations based on their purpose and influence. The new dataset, called the ACT dataset is the largest multi-disciplinary dataset of its type in existence with annotations for 11,233 citations annotated by 883 authors [Pride and Knoth, 2020].

Table 1 illustrates a sample instance from the ACT dataset. Each citation context in the dataset contains the label, "#AUTHOR_TAG", which represents the citation that is being considered. The citing_title and citing_author corresponds to the details of the document with the citation context. The dataset also has information about the cited paper (tile and author details) corresponding to the #AUTHOR_TAG. The citation_class_label represents the purpose category and the citation_influence_label corresponds to the binary class based on how influential the citation is.

The participants were provided with a labeled training dataset in the csv format with 3,000 instances, annotated using the ACT platform. Since Kaggle InClass competitions doesn't allow hosting more than one task using the same interface, separate competitions had to be created. Also, we had to split the dataset into two, based on the citation class label and the citation influence label. We also converted the categorical labels to numeric values. The citation class labels corresponds to values between 0 and 5, where each value represents the following categories:

0 - BACKGROUND

1 - COMPARES_CONTRASTS

2 - EXTENSION

3 - FUTURE

4 - MOTIVATION

5 - USES

Similarly, the citation influence labels were represented with values 0 or 1, as follows:

0 - INCIDENTAL

1 - INFLUENTIAL

Figure 1 illustrates the data distribution for Subtask A. The dataset is highly imbalanced with nearly 55% of the instances belonging to the BACKGROUND class in the training set. The FUTURE class has the lowest number of instances with just 62 and 15 instances in the training and the test dataset, respectively. The number of instances of INCIDENTAL and INFLUENTIAL classes used for Subtask B is shown in Figure 2. The dataset is relatively less skewed for Subtask B, with the number of instances associated with the incidental class (1,568) being higher than the influential class (1,432) for the training set. For both tasks, we ensured that the data distribution of categories in training set to be nearly the same as the test set. Besides the ACT dataset, participants were also encouraged to use external datasets, like the ACL-ARC [Jurgens et al., 2018], which is compatible with our dataset, for training, provided, the teams mention this while describing the systems.

3.3 The Baseline

We made an initial submission based on a simple majority class prediction as a baseline entry for both subtasks. For Subtasks A and B, the majority class corresponds to the categories, BACKGROUND and INCIDENTAL, respectively. As the competition proceeded, we also made a submission based on the BERT model [Devlin et al., 2018]. We used the pre-trained model, scibert-scivocab-uncased[5], pretrained on a sample of 1.14M multi-domain papers from the Semantic Scholar [Beltagy et al., 2019]. The 3,000 training instances were then used for fine-tuning, to obtain the task-specific results. The rational here has been to test how a state-of-the-art method, recently reported in [Cohan et al., 2019] performs compared to the methods submitted by the participants.

4 Evaluation

The evaluation was based on the test set of 1,000 examples. The test dataset was partitioned into public and private sets in Kaggle. 50% of the test set was used for the initial evaluation, and the evaluation results against it appeared on the public leaderboard as the competition progressed. The rest of the data, which is the private partition on the test file, was used for the final scoring. The private leaderboard was visible only to the shared task organisers during the competition period.

We used macro f-score for evaluating the submissions.

$$F1 - macro = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times_{Pi} \times_{Ri}}{_{Pi}+_{Ri}} \quad (1)$$

where $_{Pi}$ and $_{Ri}$ denotes the precision and recall for class i and n represents the number of classes. We chose macro f-score in light of the disproportionate distribution of output categories in our dataset and to encourage the task participants to focus on the

[5] https://github.com/allenai/scibert

Team Name	Run ID	Leaderboard	
		Public	Private
UFMG	5	**0.21460**	**0.20560**
scibert		0.17966	0.19026
Scubed	3	0.17599	0.18146
Amrita_CEN_NLP	2	0.11981	0.12542
majority_class_baseline		0.12047	0.11489

Table 2: Public and private leaderboard macro f1-scores for citation context classification based on purpose (Subtask A)

Team Name	Run ID	Leaderboard	
		Public	Private
Paul Larmuseau	1	0.57556	**0.55565**
Scubed	3	**0.59108**	0.55204
UFMG	1	**0.59108**	0.54747
Amrita_CEN_NLP	2	0.48937	0.51534
scibert		0.54747	0.50012
majority_class_baseline		0.30458	0.32249

Table 3: Public and private leaderboard macro f1-scores for citation context classification based on influence (Subtask B)

detection of the minority classes, which are particularly crucial for advancing the field of research metrics beyond just counting citations.

The submission file, in csv format, contains the unique id followed by the citation class label for Subtask A or citation influence label for Subtask B. We encouraged team submissions in Kaggle and did not set any restrictions on the team size. The limit on the number of submissions per day was set to 20. All teams were allowed to submit a maximum of 5 runs to the competition for the final evaluation for each of the tasks. The best submitted system will be used by kaggle for final scoring on the private leaderboard.

5 Participating System Description

This section presents the overview of the systems used by the participated teams, UFMG, Paul Larmuseau, Scubed and Amrita_CEN_NLP in the 3C shared task. Except for Paul Larmuseau, rest of the teams participated in both the tasks. The teams that participated in both tasks used the same approach while making submissions to Subtask A and Subtask B.

5.1 UFMG

Team UFMG[6] explores the possibility of enhancing the results by using a combined text representations for capturing the statistical, topical and the contextual information. For this, they chose Term Frequency-Inverse Document Frequency (TF.IDF) for word representation (upto bigrams), Latent Dirichlet Allocation (LDA) for topic extraction from citation context and finally GloVe embeddings[7] to obtain the word vector representation for capturing the word co-occurrences. The team obtained the highest score of 0.2056 for subtask A by combining the above mentioned word representations for the passive aggressive classifier, an incremental learning mechanism. However, for Subtask B, UFMG obtained the best overall score of 0.54747, finishing as third on the leaderboard, just by using a single feature, TF.IDF. Furthermore, by using additional feature like self citation along with the TF.IDF, the team claims to have obtained a 3.1 % improvement in the final score for Subtask B [Valiense de Andrade and Goncalvesh, 2020].

5.2 Scubed

The team Scubed[8] applied TF.IDF on the columns, citing title, cited title and the citation context in the dataset. They used off-the-shelf machine learning based models, including Logistic Regression (LR), Random Forest (RF), Gradient Boosting Classifier (GBT) and two variants of the Multi-Layer Perceptron (MLP) classifiers. For Subtask A, the best performing model using MLP obtained a private score of 0.18146 and the team finished third. However, for the binary classification task, RF achieved the best score and the team finished second on the leaderboard with a macro f-score of 0.55204. The team also reports a per category model evaluation using the truth labels of the test set [Mishra and Mishra, 2020a,b].

5.3 Paul Larmuseau

The best system in the subtask B was that of Paul Larmuseau[9]. The team used a combined TF.IDF weighting and fasttext embedding, consisting of 1 million word vectors trained on Wikipedia 2017[10]. Another important feature used by the team was

[6] 10.6084/m9.figshare.12638807

[7] https://nlp.stanford.edu/projects/glove/

[8] 10.6084/m9.figshare.12638846

[9] 10.6084/m9.figshare.12638840

[10] https://fasttext.cc/docs/en/english-vectors.html

the cosine similarity, calculated between the citing title and a combination of cited title and the citation context. As part of the pre-processing step, they also experimented with feature scaling (based on the maximum absolute values) and dimensionality reduction (single value decomposition regression) techniques. The team experimented with different approaches and obtained the highest private score of 0.55566 using LR, finishing first in Subtask B [Larmuseau, 2020].

5.4 Amrita_CEN_NLP

The team Amrita_CEN_NLP[11] used Word2Vec for extracting the contextual information and feature representation. In order to built the vocabulary, the team used the shared task training and the test dataset. The team experimented with different classifiers like LR, Decision Tree (DT), k-Nearest Neighbour (k-NN), LR and Ada Boost. A cost sensitive learning approach for assigning separate weights was used for Subtask A, to address the class imbalance issue. The best score for both subtasks was achieved using RF [B and K.P, 2020].

6 Results

Table 2 shows the public and the private macro f-scores obtained by the teams for Subtask A. The highest public and private macro f-score was obtained by the team, UFMG. The submission based on scibert model scored the second best result with a private score of 0.19026. This was followed by the teams scubed and Amrita_CEN NLP in the third and fourth positions. All the teams substantially outperformed the majority class baseline classifier. Since the dataset for purpose classification task was highly skewed, with the majority of the classes belonging to the BACKGROUND class and the fact that we used macro f-score for evaluating the systems, all the systems submitted for this task scored less when compared to the Subtask B.

The results for the final evaluation of systems submitted for Subtask B is shown in Table 3. The highest performing system, submitted by Paul Larmuseau achieved a private macro f score of 0.55565, ranking as first for Subtask B. However, two other systems submitted by the teams Scubed and UFMG obtained an even higher score of 0.59108 on the public data. The deep learning based language model scibert achieved lesser score compared to the rest of the submissions using simpler machine learning model for this binary classification task. Not surprisingly, the systems submitted to Subtask B achieved better results when compared to the other task, because of the lesser number of categories and less skewness in the data distribution.

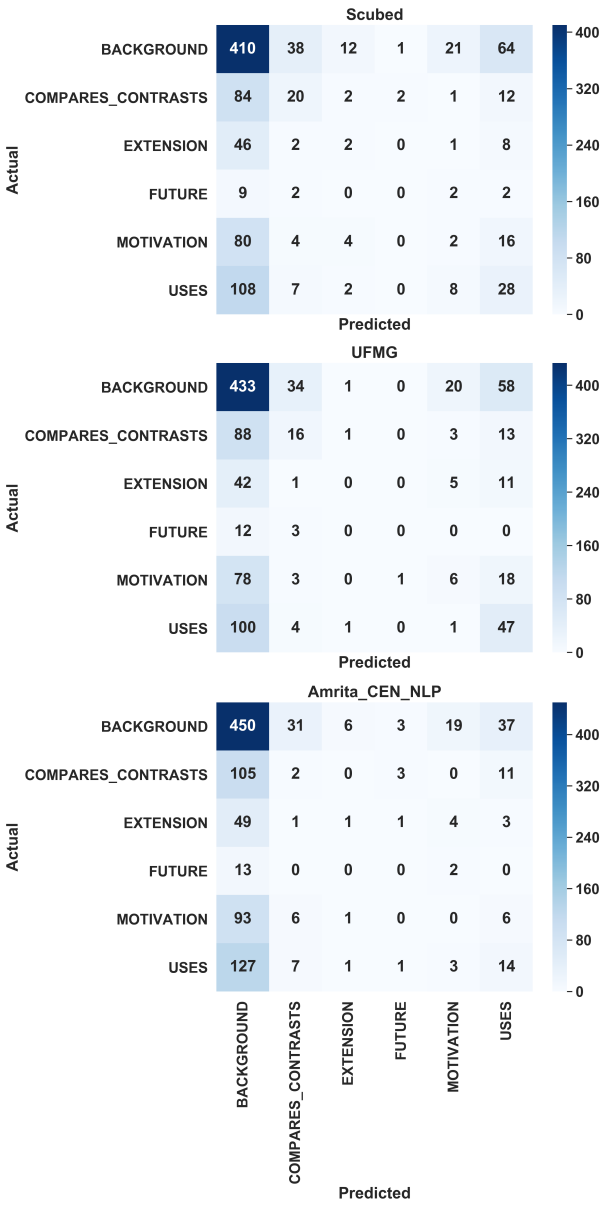

Figure 3: Confusion Matrix for subtask A

7 Discussion

The 3C Shared task is the first open competition for citation context classification. This shared task could be considered as a new benchmark for these tasks as we release both the data and the source code of all the submitted systems. All the teams that participated in this shared task used simple

[11] 10.6084/m9.figshare.12638849

Run ID	Team	Field Used	Model	Features	Public Score	Private Score
1	UFMG	citation_context	Passive Aggressive	TF.IDF	0.19829	0.19425
2				LDA	0.12923	0.15826
3				GloVe	0.12047	0.11489
4				TF.IDF+LDA	0.19124	0.19572
5				TF.IDF+GloVe	0.19945	0.20037
6				TF.IDF+LDA+GloVe	**0.21460**	**0.20560**
1	Scubed	citing_title, cited_title, citation_context	GBT	TF.IDF	0.15001	0.14381
2			RF	TF.IDF	0.14262	0.15826
3			MLPC	TF.IDF	**0.17599**	**0.18146**
1	Amrita_CEN_NLP	citation_context	DT	Word2Vec	**0.20709***	0.16732*
2			RF	Word2Vec	0.11981	0.12542
3			kNN	Word2Vec	0.16623*	0.13563*
4			Adaboost	Word2Vec	0.12047*	0.11489*
5			LR	Word2Vec	0.17309*	**0.19530***

* Post-Evaluation Results

Table 4: Overall Result (Subtask A)

Run ID	Team	Field Used	Model	Features	Public Score	Private Score
1	Paul Larmuseau	cited_title, citing_title, citation_context	LR	TF.IDF	**0.57556**	0.55565
2			LR	fasttext + TF.IDF	0.54726 *	**0.60333** *
1	UFMG	citation_context	Passive Aggressive	TF.IDF	**0.59108**	0.54747
2				LDA	0.30458	0.32249
3				GloVe	0.30458	0.32249
4				TF.IDF+LDA	0.32707	0.36156
5				TF.IDF+GloVe	0.30458	0.32249
6				TF.IDF+LDA+GloVe	0.30458	0.32249
7				TF.IDF+self_citation	0.57556 *	**0.55565** *
1	Scubed	citing_title, cited_title, citation_context	LR	TF.IDF	0.30458	0.32249
2			GBT	TF.IDF	0.56473	0.52351
3			RF	TF.IDF	**0.59108**	**0.55204**
4			MLP-3	TF.IDF	0.51589	0.48187
1	Amrita_CEN_NLP	citation_context	DT	Word2Vec	0.47565	0.47596
2			RF	Word2Vec	**0.48937**	**0.51534**
3			kNN	Word2Vec	0.46386	0.43769
4			Adaboost	Word2Vec	0.30458	0.32249
5			LR	Word2Vec	0.31250	0.32579

* Post-Evaluation Results

Table 5: Overall Result (Subtask B)

machine learning-based classifiers, including logistic regression, random forest, and multi-layer perceptron. One of the teams experimented with the online learning technique for faster computation. As with feature representation, the conventional approach used by the majority of the teams was TF.IDF. The prospect of employing word vectors developed using Wikipedia, the shared task dataset and the use of pre-trained embeddings like GloVe were explored by the teams.

Figure 3 shows the confusion matrix for the best systems submitted by the teams Scubed, UFMG, and Amrita_CEN_NLP for the subtask A. The most successfully classified category is BACKGROUND. The winning team, UFMG, classified nearly 80% of the BACKGROUND class instances correctly. The number of true positives for the minority class FUTURE is zero, which implies that none of the above mentioned teams could successfully categorise the instances to this class. The imbalanced nature of the subtask A dataset significantly affects the performance of the systems submitted by teams, which is one of the challenging aspects as far as citation function classification task is concerned.

Tables 4 and 5 displays the public and private scores obtained by teams for the different systems they submitted for subtask A and subtask B respectively. All the teams for both tasks used the data field, citation_context as the main source of semantic information for feature extraction, and classification. Two teams also examined citing_title and the cited_title fields for extracting useful features. Since Kaggle allows late submissions for the hosted competitions, the participants can still submit results to get better scores, although this will not be visible on the public and the private leaderboard. Both the tables also contain the post-evaluation results obtained by some of the teams.

The current deep learning based state-of-the-art language models like scibert could not achieve better results on our dataset, and as the leaderboard indicates, such sophisticated models are beaten by more simpler methods, that are significantly less computationally expensive on this task. One possible reason for this could be the lesser number of training instances we provided to the participants.

8 Conclusion

Citations, which act as a connection between the cited and the citing articles, cannot be treated equally and serve different purposes. Traditional citation analysis based on mere citation counts take into consideration just the quantitative factors. Analysing the citation context for classifying citations based on their function and influence has many applications and the most important being its implementation in the research quality evaluation. One of the greatest challenges faced in the citation context analysis for identifying the citation function and its influence is the absence of multi-disciplinary datasets and unavailability of medium to fine grained schemes which sufficiently captures information for citation classification [Hernández-Alvarez and Gómez, 2015]. Although previous works on the problem of citation context classification exist, lack of shared datasets, common conventions and annotation schemes caused the benchmarking of systems on the same tasks difficult.

The 3C Shared task constitutes the first systematic effort to a) compare different methods on the same data, b) on the same classification taxonomy across two previously reported tasks, and c) on multi-disciplinary data. We propose the unifying framework of the 3C shared task to be used as a standardised benchmark for this task, as we make all the submitted systems to this shared task, publicly available. We believe this will allow future comparison of participating systems head-to-head on the same data and task. The results obtained by the teams indicate the relevance of the simple machine learning based models over complex deep learning based approaches. The winning team for the subtask A, UFMG obtained an overall score of 0.19425. The team, Paul Larmuseau finished at first position on the leaderboard with a macro f score of 0.55565 for subtask B.

References

Premjith B and Soman K.P. Amrita_cen_nlp wosp 3c citation context classification task. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China, 2020.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Myriam Hernández-Alvarez and José M Gómez. Citation impact categorization: for scientific literature. In *2015 IEEE 18th International Conference on Computational Science and Engineering*, pages 307–313. IEEE, 2015.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

Norman Kaplan. The norms of citation behavior: Prolegomena to the footnote. *American documentation*, 16(3):179–184, 1965.

Petr Knoth and Zdenek Zdrahal. Core: three access levels to underpin open access. *D-Lib Magazine*, 18 (11/12):1–13, 2012.

Paul Larmuseau. Find influential articles in a dataset. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China, 2020.

Shubhanshu Mishra and Sudhanshu Mishra. Scubed at 3c task a - a simple baseline for citation context purpose classification. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China, 2020a.

Shubhanshu Mishra and Sudhanshu Mishra. Scubed at 3c task b - a simple baseline for citation context influence classification. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China, 2020b.

Michael J Moravcsik and Poovanalingam Murugesan. Some results on the function and quality of citations. *Social studies of science*, 5(1):86–92, 1975.

David Pride and Petr Knoth. Incidental or influential?– a decade of using text-mining for citation function classification. 2017.

David Pride and Petr Knoth. An authoritative approach to citation classification. In *2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Virtual - China, 2020.

David Pride, Petr Knoth, and Jozef Harag. Act: an annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 329–330, Urbana-Champaign, Illinois, 2019. IEEE.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110, 2006.

Claudio Moises Valiense de Andrade and Marcos Andrer Goncalvesh. Combining representations for effective citation classification. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China, 2020.